Exhibit 174

High Correlations Between Predicted and Actual Ballots Do Not Imply Fraud

Justin Grimmer*

Matthew Tyler†

June 2021

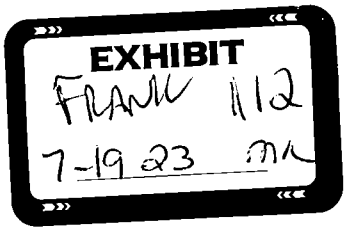
1 Introduction

Douglas Frank asserts that a strong correlation between the predicted count of votes from age groups and the actual count of votes across counties in swing states is evidence that a "key" or "algorithm" was used to determine the vote before the election. In this brief report, we show that there is a high correlation between predicted and actual ballots across all states where we have adequate data to make the assessment. There is nothing remarkable about this high correlation. Because the number of registrants appears in both terms of the correlation Frank's correlation is artificially inflated. If we make the same effective comparison without including the number of registrants, then we find more modest correlations. Additionally, we show that Frank's high correlations are found even if the predicted turnout rates are generated with substantial noise.

2 Procedure and Replication

Frank never responded to our request for replication code or data, so we used his public descriptions to replicate his results. Specifically, using the L2 data we calculated the turnout rate for

*Democracy and Polarization Lab and Hoover Institution, Stanford University
†Democracy and Polarization Lab, Stanford University



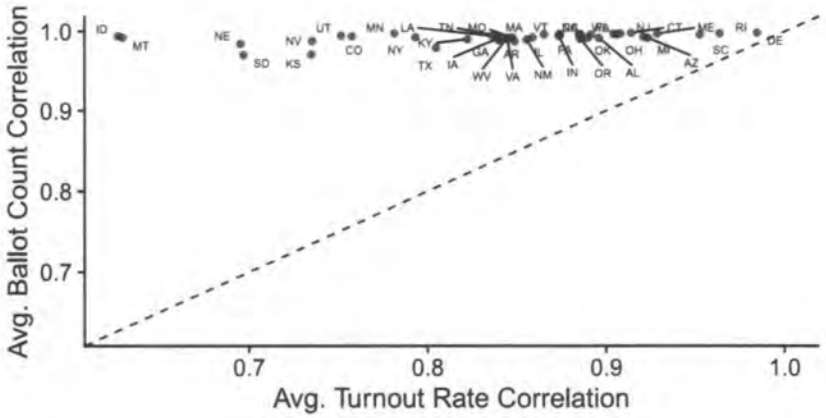


Figure 1: High correlations between actual and predicted ballots in counties is found in each state.

registered voters for each age from 18-100 for all states where we observed a birthdate for more than 90% of registrants.[1] With the turnout rate calculated at the state level for each age, we then regressed these turnout rates on a six-degree polynomial to create a predicted turnout rate for each age. We then applied the predicted turnout rate to each county in each state. Finally, following Frank's analysis, we calculated the correlation in each county between the actual count of ballots and the predicted count of ballots by age. The predicted count of ballots was obtained by multiplying the number of registrants in that age by the predicted turnout rate for that age. As Figure A.1 in the appendix shows, we are able to replicate the shape from Frank's original analysis, though it is unclear why Frank's "key" goes above 1 (an impossible turnout rate).

3 High Correlations Are Found in Every State

In Figure 1 we present the results of applying Frank's analysis to each state where we have adequate data. The vertical axis presents the statewide average correlation coefficient (what Frank denotes by *R*) between the actual count of ballots and the predicted count of ballots within each county. We find an impressively high correlation in each of the states, not just those states that Frank analyses. In fact, across all states included in our analysis the average correlation is 0.992.

Therefore, there is nothing special about the states that Frank analyzes. It doesn't appear that

[1]The excluded states are AK, DC, HI, MD, MS, ND, NH, WI, WY

states who make use of Dominion machines have particularly high correlations, nor is it the case that Trump's performance in a state helps predict the correlation.

The high correlation between the predicted and actual count of ballots is also not found when assessing the correlation between the predicted and actual turnout rates. While Frank appears to imply that turnout rates are effectively constant across counties, this is decidedly not the case. The horizontal axis of Figure 1 computes the statewide average correlation between the predicted turnout rate in counties and the actual county turnout rates. These are noticeably smaller, with statewide average correlations ranging from approximately 0.6 to 0.9. The distribution of county rate correlations is shown in Figure A.2, which shows an average correlation of only 0.76. Moreover, the average ballot count correlation (the vertical axis) appears largely unrelated to the turnout rate correlation — which suggests that the ballot count correlation is measuring something else.

4 Correlating a Variable With Itself

The high correlation between actual and predicted ballots by age that Frank reports is in fact the result of effectively correlating a variable with itself. Frank computes the correlation:

$$cor(\text{Actual Count of Ballots}, \text{Predicted Count of Ballots}) =$$
$$cor(\text{Turnout} \times \text{Number of Registrants}, \text{Predicted Turnout} \times \text{Number of Registrants})$$

The *Number of Registrants* term appears on both sides of the correlation. This will lead to an artificially higher correlation under a wide range of conditions (especially here since the predicted and actual turnout rates are confined to a narrow range). To provide intuition for this empirically, in Figure 2 we show that random variation in predicted turnout rates across counties by age would also produce the high correlation between predicted and actual ballot counts that Frank reports. To produce this Figure, we computed random turnout rates in each county, with the amount of

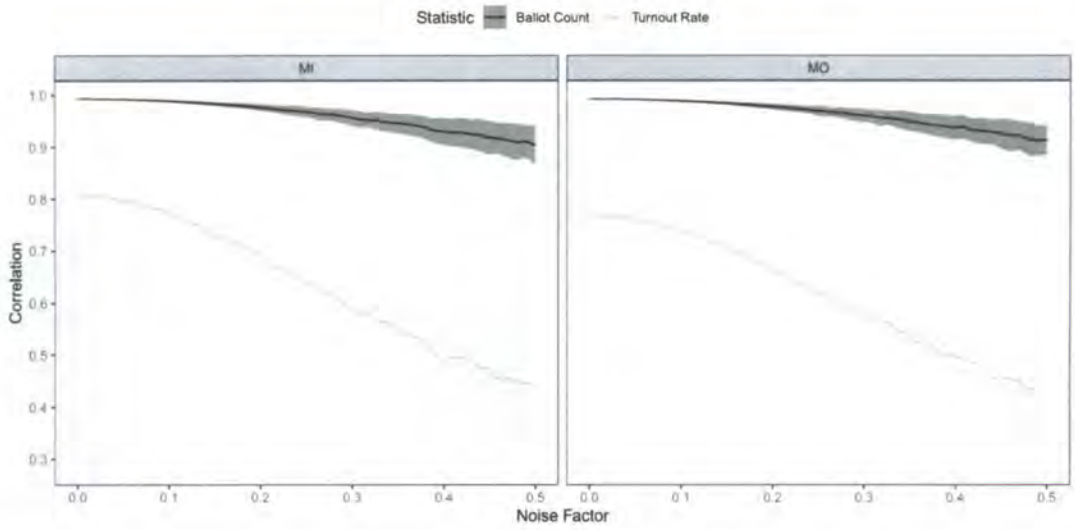


Figure 2: The high correlation between predicted and actual ballot counts remains even when noise is added to the predictions.

"noise" added to the true (state-level) turnout rate given on the horizontal axis.[2] We see that for a wide range of noise factors there is an impressively high correlation between actual and predicted ballot counts even as the correlation between predicted and actual turnout rates falls. This remains true even as considerable noise is added to the simulation. And this is true in a swing state (Michigan, left-hand plot) and a state Trump won handily (Missouri, right-hand plot).

5 Conclusion

Frank's analysis is not evidence of voter fraud. Instead, we show that the high level of correlation that Frank finds is largely the result of correlating a variable with itself. In fact, the artificially high correlations that Frank reports remain high even when random noise is added to the predicted turnout rates. Therefore, it is no surprise that all states with available age data — not just swing

[2]In particular, the predicted turnout rate with noise is $\Phi(\Phi^{-1}(\text{Original Prediction}) + (\text{Noise Factor}) \times Z)$ where Φ is the Gaussian cumulative distribution function (CDF) and Z is a standard Gaussian variate. This is equivalent to adding Gaussian noise to the original prediction on the Gaussian latent scale. To estimate the expected value of this process with Monte Carlo, we repeat this 100 times for each county and noise factor and report the average plus or minus two Monte Carlo standard errors.

states or states accused of voter fraud – show similarly high but unsurprising correlations.

A Appendix

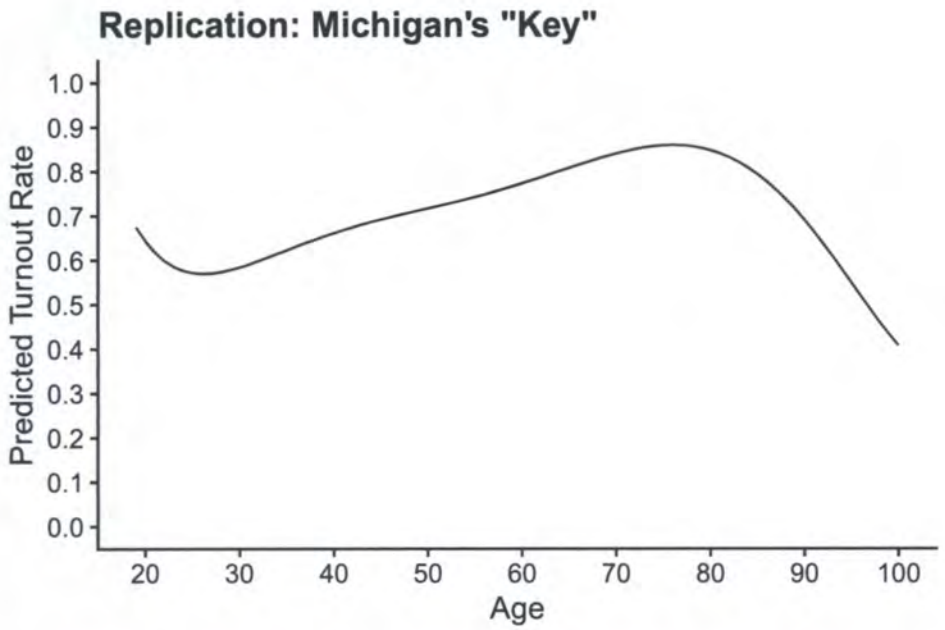


Figure A.1: Predicted turnout (from a 6th-order polynomial) by age.

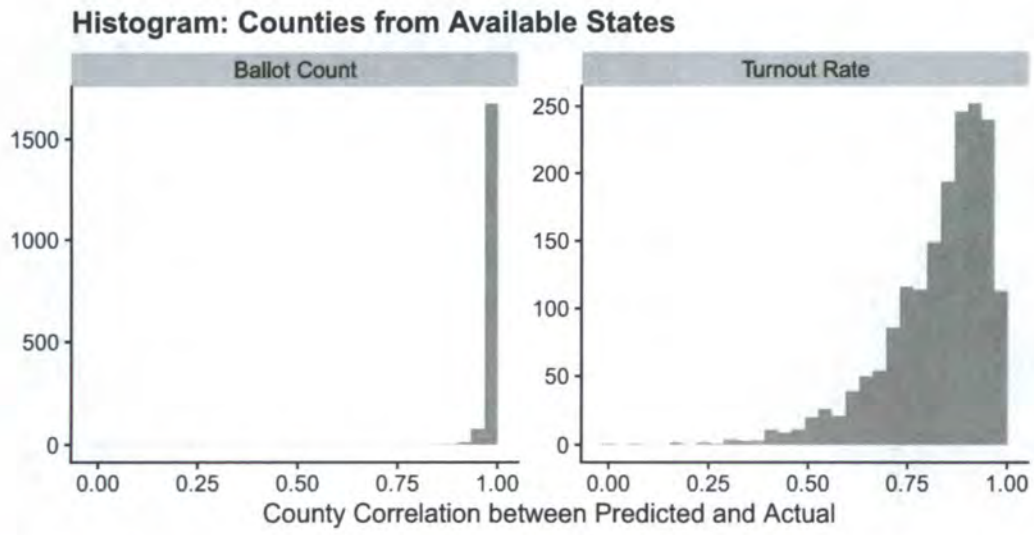


Figure A.2: Histograms of county correlations between predicted and actual ballot counts (left) and turnout rates (right).